Chapter 4: The Case of Old English Word Order

4.1 Filters on Data Intake for Syntactic Learning

The phenomenon I examine in this chapter is an instance of syntactic learning, specifically the alternation between Object-Verb (OV) and Verb-Object (VO) order in Old English. This case is another example where the learner has two hypotheses under consideration. However, unlike the case of anaphoric *one*, the final state for *adults* in Old English is argued to be probabilistically distributed between the two hypotheses (Pintzuk, 2002; Kroch & Taylor, 1997; Bock & Kroch, 1989). Evidence for this mixed adult state comes from texts in which both alternates are exhibited by a single author. This is in contrast to final state where only one hypothesis is accessed (i.e. only one structural rule used) by adults.

The hypothesis space for Old English OV/VO order consists of two hypotheses that overlap, but do not have a subset-superset relation. Both the OV and VO hypotheses have data that will be unambiguous. In addition, there is a quantity of data that is ambiguous between the two hypotheses since it can be analyzed successfully given either hypothesis. The updating procedure is based off the one described in the mathematical framework in chapter 2. I then use this definition of the hypothesis space and the updating procedure to investigate two filters on data intake proposed for syntactic learning.

The two filters in question bias learners away from potentially misleading ambiguous data in the input, both stemming from a presumed preference for "simple" data (Dresher, 1999; Lightfoot, 1999, 1991; Fodor, 1998a). These filters use a structurally-based notion of simplicity. The first claims that children learn only from unambiguous data (Dresher, 1999; Lightfoot, 1999; Fodor, 1998a), and consequently do not activate the update algorithm whenever data is perceived as ambiguous. The second proposal restricts learning to the data points found in "simple" clauses (Lightfoot, 1991), where simple clauses are defined as matrix clauses. If there are available data points in embedded clauses, the update algorithm again is not activated and these data are effectively ignored by the learner.

These filters are motivated by the perceived informativity and ease of comprehensibility of the relevant data. As we saw in the previous chapter, an unambiguous data point allows the learner to be maximally confident in whichever hypothesis the data point signals. So, the most probability is shifted when the learner encounters an unambiguous data point. We can view this as unambiguous data points being the most informative data points available to the learner. For simple clauses, it has been claimed that children might restrict their attention to simple, subparts of utterances (Morgan, 1986), perhaps because of general cognitive restrictions on the complexity of data that they can handle. So, matrix clauses, being "simpler", are arguably easier for learners to extract information from.

Nonetheless, filtering the data is not without its drawbacks. The filters proposed above will radically truncate the data intake set. It is well known that sparse data can inhibit a probabilistic model's ability to converge on a solution. Thus, we must determine if the subset of data circumscribed by these two filters can still allow learning to succeed, even if the subset is significantly smaller than the input data set.

In Old English, as we have already noted, the adult state is a probabilistic distribution between the two hypotheses, OV and VO word order. Because the target state is *not* an endpoint (either all OV or all VO word order), it is more difficult to gauge learning success. How close does the learner have to get to the adult probability distribution in order for learning to be deemed successful?

At this point, we can make use of the fact that languages change over time. Specifically in the case of Old English, the population shifts from an OV-biased distribution around 1000 A.D. to a VO-biased distribution around 1200 A.D. (YCOE Corpus, Taylor et al., 2003; PPCME2 Corpus, Kroch & Taylor, 2000). It has been proposed that certain types of change (such as the shift in Old English) result from a misalignment of the child's hypothesis and the adult's analysis of the same data (Lightfoot, 1999; 1991). In other words, language change in this case results from *imperfect* learning of a very particular kind.

Specifically, the idea is that language change in this case occurs because learners misconverge on the probability distribution; the learner's probability distribution is very slightly different from the adult's probability distribution. The key point is that the amount of difference between the learner's probability distribution and the adult's probability distribution will influence the rate of language change in a population over time. In order to model change at an attested pace, the acquisition model must hypothesize exactly the right amount of difference between the learner's and adult's probability distributions.

Therefore, "successful" learning is defined as learning that leads to exactly the right amount of *misconvergence* within the individual learner. This amount of misconvergence within the individual then leads to language change over time within the population of individuals. We will find that the amount of misconvergence depends greatly on how the input is filtered during learning. Thus, we can test proposals about data filtering by using models of language change.

It is important to note the correlation between successful learning and imperfect learning for certain cases of language change. Often, language learning research in synchronic cases may focus so much on the learner's ability to reach the target adult state that we may overlook the fact that perfect learning will not necessarily lead to success in diachronic cases. This is because perfect learning would entail no change over time. This then creates a certain tension on the demands of a successful learning model – it must be good enough that learners can communicate effectively with the remainder of the population, but not so good that language change is impossible. So, using successful language change as a metric for successful language learning attempts to keep this second point in mind.

We will find, perhaps surprisingly, that the two proposed filters on data intake are crucial for a successful model of Old English language change that describes a population which begins strongly OV-biased at 1000 A.D. and ends strongly VO-biased at 1200 A.D. Without these filters, the simulated learners are unable to misconverge the precise amount necessary for the modeled population's rate of change to match the historically attested population's rate of change. This supports the existence of these two filters on data intake during the normal course of syntactic learning.

The chapter proceeds as follows. First, I will discuss the two filtering proposals in detail. Then, I will examine the available information on the language change in Old English. After that, I will discuss the model of language learning and language change that I will use. Finally, I will present the modeling results and discuss their implications for language learning.

4.2 Restricting the Data Intake

4.2.1 Unambiguous Data

4.2.1.1 Unambiguous Data for OV and VO Word Order

Unambiguous data is defined within a hypothesis space of opposing analyses for a certain piece of linguistic structure, such as OV or VO word order. Ambiguity is often faced by a child choosing the correct grammar for his or her language. Let's consider a simple example. The child has to decide whether the stream of encountered speech belongs to a VO (Verb before Objects) language requiring rules like (1) or to an OV (Objects before Verbs) language requiring rules like (2).

(1) VO rule set examples
(a) VP → V NP PP (b) VP → V NP

(2) OV rule set examples
(a) VP → NP PP V (b) VP → NP V

Modern English chooses the VO rule set (1). Modern Dutch and German choose the OV rule set, which includes those in (2). However, modern Dutch and German also generate strings that are compatible with some of the rules in set (1), such as in (3) below:

(3) Ich$_{\text{Subj}}$ sehe$_{\text{TensedVerb}}$ [den Fuchs]$_{\text{Obj}}$
I see the fox
'I see the fox.'

This example demonstrates an option available in modern Dutch and German which moves the tensed Verb of the matrix clause to the "second" phrasal position in the matrix clause, known as V2 movement (Lightfoot, 1999; Kroch & Taylor, 1997; among many others). The tensed Verb *sehe* moves from its original position (after *den Fuchs*) to the second phrasal position in the sentence, and some other phrase (*Ich*) moves to the first phrasal position, as in (4).

(4) Ich$_{\text{Subj}}$ sehe$_{\text{TensedVerb}}$ t_{Subj} [den Fuchs]$_{\text{Obj}}$ $t_{TensedVerb}$.
I see the fox
'I see the fox.'

Given the example in (3), one might reasonably wonder why we posit the analysis in (4) instead of simply assuming that modern German (and Dutch) word order is VO. The reason is that VO order does not appear in matrix clauses across the board. Languages like modern Dutch and German use VO order only for tensed Verbs in matrix clauses. Non-tensed Verbs in matrix clauses and all Verbs in embedded clauses obey OV order and appear after the Object. This forces us to assume a basic OV word order with an additional operation that moves the tensed Verb in matrix clauses.

In the bold part of (5a), we see the basic OV order appearing in the embedded clause as *den Fuchs sehen kann* (Object Non-TensedVerb TensedVerb). In (5b), the non-tensed Verb *sehen* appears in the matrix clause after the Object *den Fuchs*, again displaying the OV order. The V2 rule moves the tensed modal *kann* to the second phrasal position, and the Subject *Ich* moves to the first phrasal position.

(5a)	Ich$_{\text{Subj}}$	denke$_{\text{TensedVerb}}$,	das	ich	**[den Fuchs]$_{\textbf{Obj}}$**
	I	*think*	*that*	*I*	*the fox*

	sehen$_{\textbf{Non-TensedVerb}}$	**kann$_{\textbf{TensedVerb}}$**
	see	*can*

‘I think that I can see the fox.’

(5b)	Ich$_{\text{Subj}}$	kann$_{\text{TensedVerb}}$	t_{Subj}	[den Fuchs]$_{\text{Obj}}$	sehen$_{\text{Non-TensedVerb}}$	$t_{TensedVerb}$
	I	*can*		*the fox*	*see*	

‘I can see the fox.’

At the beginning of language learning however, the child has not set the word order parameter for the language. Therefore, both the OV and VO hypotheses are available with some probability. The matrix clause *Ich sehe den Fuchs* can be covered by both hypotheses. The OV hypothesis can use the analysis described in (4), matrix OV order with the V2 movement rule; the VO hypothesis can use the analysis in (6), matrix clause VO order without the V2 movement rule (which is the analysis used for modern English).

(6)	Ich$_{\text{Subj}}$	sehe$_{\text{TensedVerb}}$	[den Fuchs]$_{\text{Obj}}$.
	I	*see*	*the fox*

‘I see the fox.’

Data points like (3) are therefore ambiguous between the two hypotheses under consideration. A proposal to filter data intake down to the unambiguous data points would cause the learner not to activate the update procedure when encountering ambiguous data points.[20] Instead, the learner uses only data points perceived as unambiguous. Examples of perceived unambiguous data are in (5)

[20] Otherwise, the learner would require some strategy for how to update the probabilities when encountering ambiguous data, as we saw in the previous chapter.

above. In (5a), if the child uses embedded clause data as intake, then the presence of the Verbs (both tensed *kann* and non-tensed *sehen*) after the Object would signal that the VO hypothesis is correct. In (5b), the presence of the non-tensed Verb *sehen* after the Object again implicates the OV hypothesis since that order would not be generated by a VO system.

4.2.2.2 Identifying Unambiguous Data

If we believe that children filter their intake for syntactic learning down to unambiguous data, it is important to provide a plausible method for identifying unambiguous data. Two methods have been proposed to identify unambiguous data: the domain-specific knowledge of cues (Dresher, 1999; Lightfoot, 1999) and the domain-specific procedure of parsing (Fodor, 1998; Sakas & Fodor, 2001).

A cue for identifying unambiguous data is defined as a specific configuration in the surface structure of the data point that signals one parameter value (hypothesis) is correct. The knowledge of what a cue for a given parameter value looks like is often presumed to already be available to the learner (Dresher, 1999; Lightfoot, 1999), whether innately specified or derived through some other knowledge. A cue for OV/VO word order proposed by Lightfoot (1999) is described in (7).

(7) The Object is adjacent to the Verb (on the appropriate side) and the Verb is *not* in the second phrasal position.

This is considered a cue because the V2 movement rule deriving a VO order from an underlying OV order only allows a single phrasal constituent to come before the Verb. If the Verb is preceded by more than one phrasal constituent, then its position is not the result of V2 movement.[21] The form of this cue could be an underspecified piece of sentence structure (figure 28 below) or simply a linear pattern retrievable from the observable data (8). Both are representations of the domain-specific knowledge that a cue describes.

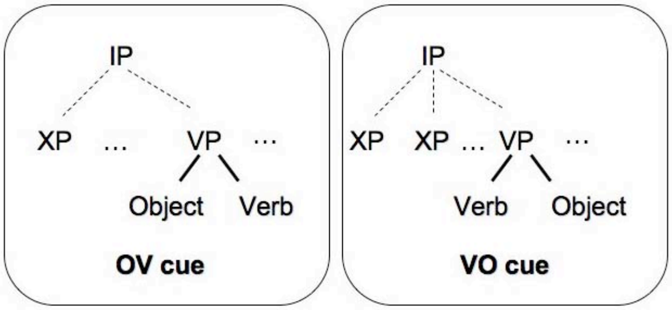

Figure 28. Underspecified pieces of sentence structure that could be the learner's representation of a cue for OV vs. VO word order, as described by Lightfoot (1999).

[21] Note that this is the learner's *perception* of the data, given a restricted knowledge base. The adult grammar, in actuality, may contain other grammatical rules that allow V2 movement to create a clause with the Verb in the third position. Thus, the learner may perceive data as "unambiguous" that is ambiguous when a fuller range of grammatical rules is considered.

(8) Linear patterns that could be the learner's representation of a cue for OV vs. VO word order.

(a) OV cue: $[\ ]_{XP}$ … Object Verb …
(b) VO cue: $[\ ]_{XP1}$ $[\ ]_{XP2}$ … Verb Object …

To identify unambiguous data, the learner matches the data point (or relevant piece of the data point) to the cue. Example sentences that would match these cues are in (9).

(9a) Matching OV cue: Subject Object Verb.
Ich denke, das *ich den Fuchs sehe.*
(XP = Subject, … = *null*)
(9b) Matching VO cue: Adverb Subject Verb Object.
Yesterday, I saw a dragon.
(XP1 = Adverb, XP2 = Subject, … = *null*)

The cues method gives sentences like these privileged status, and such sentences are viewed as unambiguous evidence for the associated parameter value, OV or VO.

An alternative approach is to use the learner's natural language comprehension processes to discover if a data point should be considered unambiguous for OV/VO order (Fodor, 1998b; Sakas & Fodor, 2001). The learner assigns possible structures to (or *parses*) the datum with all values of the *relevant* parameter set (in this example, the relevant parameter set *PS* = {OV/VO, +V2/-V2}).[22] If only one value of a parameter (e.g. OV) will allow a successful parse of the entire data point, then that data point is classified as unambiguous for that value of that parameter. This procedure is shown in (10).

(10) Parsing to identify unambiguous data for basic word order using the set of parameter values PS = {OV/VO, +V2/-V2}

(a) Data point: *Subject Object Verb.*
Sets of values from PS that will lead to a successful parse of the data =
{OV, -V2}
In this case, the only combination of values that will allow a successful parse is OV and –V2. Therefore, given this set of relevant parameter values, this data point is unambiguous for both OV and –V2.

(b) Data point: *Subject TensedVerb NonTensedVerb Object.*
Sets of values from PS that will lead to a successful parse of the data =
{VO, +V2}, {VO, -V2}
In this case, two combinations of values will allow a successful parse of the data point, and both use the VO value (and neither use the OV value). Either value of the V2 parameter can be used in combination with the VO value,

[22] Note that the relevant parameter set for the learner may be (and likely is) a subset of the entire adult parameter set.

however. Therefore, given this set of relevant parameter values, this data point is unambiguous for VO only.

(c) Data point: *Subject Verb Object.*
Sets of values from PS that will lead to a successful parse of the data =
{OV, +V2}, {VO, -V2}, {VO, +V2}
In this case, three combinations of values will allow a successful parse of the data point. Importantly, neither parameter value for either parameter is crucial for parsing success. There is at least one combination that uses the OV value, at least one that uses the VO value, at least one that uses the +V2 value, and at least one that uses the –V2 value. Therefore, given this set of relevant parameter value, this data point is *not* unambiguous for any values of any parameters.

We will return in the next chapter to the discussion of the benefits and drawbacks of each method that the learner could use in identifying unambiguous data. For the case of Old English OV/VO order discussed in this chapter, both methods will identify the same set of utterances as unambiguous data, provided the relevant parameter set for parsing is restricted as described above.[23]

4.2.2.3 Unambiguous Data Summary

The unambiguous data filter reflects a very simple idea: the child learns only from the data perceived as "clean", instead of guessing about data perceived as "unreliable". If the child is using cues, clean data are identified by the specific rubric of the cue. If the child is using parsing, clean data are identified by having only one parameter value that yields successful parsing. For both methods, it is important to note that a data point is unambiguous relative to a given parameter. A data point unambiguous for parameter P1 may not be unambiguous for another parameter P2. For instance, as we saw in (10b), a data point can be unambiguous for VO order while being ambiguous for the V2 movement operation.

In addition, an unambiguous filter reduces the set of data a child can learn from (since some data in the input are classified as unambiguous). It is therefore quite important that there be enough data left in the child's intake to learn from. If the data perceived as unambiguous appear in sufficient quantity in the input, the learner will converge on the "correct" probability distribution for that parameter. Otherwise, the individual learner within the population will not be able to converge on the correct probability distribution, and will instead remain near the initial probability distribution. Once individuals are unable to converge on the correct probability distribution, language change in the population as a whole will grind to a

[23] Specifically, the relevant parameter set for parsing should not include operations that can influence the position of the Object with respect to the Verb, such as Heavy Noun Phrase shift which will move the Object to a position following the Verb if the Object is phonologically "heavy enough". If the parameter set *did* include operations like this, many more data points would be considered ambiguous and therefore unusable for a learner employing an unambiguous data filter.

halt. Thus, it is critical for the feasibility of an unambiguous data filter that the unambiguous data not be too sparse in the input.

4.2.3 Simple Clauses

The potential problem of data sparseness becomes worse when we add a proposal to learn from data in simple clauses only: the "degree-0" learning filter of Lightfoot (1991). Degree refers to the level of embeddedness. I adopt Lightfoot's terminology "degree-0" to refer to matrix clauses and "degree-1" to refer to embedded clauses.[24] This filter is motivated by a claim that it lessens the cognitive load of the learner; children use only structural information that spans a single matrix clause and at most a complementizer in the embedded clause.[25] A learner using this filter would not use data such as (5a) as evidence for the OV order of German, since the useful structural information signaling OV order is in the embedded clause. Nonetheless, examples such as (5b) that contain non-tensed Verbs adjacent to the Object in the matrix clause are still in the degree-0 learner's intake.

4.2.4 The Influence of Input Filtering on Old English Language Change

Potential data sparseness aside, filtering of the input can go a long way toward explaining how changes to a language's structure can spread fairly rapidly through a population. Filtering requires learners to learn only from a specific subpart of the observable data. If that subpart changes (perhaps due to external factors) so that it does not accurately reflect the adult probability distribution for the language as a whole, then children will "mislearn" the adult probability distribution. These children subsequently contribute observable data to the next generation of children, who will subsequently "mislearn" the previous children's "mislearned" probability distributions. This continues, spreading through the exponentially growing population[26], until the population as a whole has shifted its probability distribution dramatically.

The loss of a strongly OV distribution in Old English is an especially interesting language change because the degree-0 unambiguous data distribution of the two word orders appears to be significantly different from the average adult's

[24] Lightfoot's work follows Wexler & Culicover (1980) and Morgan (1986), who argue for less restrictive constraints on the learning domain.

[25] Note that this motivation wouldn't necessarily hold for head-final languages like Japanese where the matrix clause can be split into two parts by an embedded clause: $\text{Subject}_{\text{Main}}\ldots\text{Subject}_{\text{Embedded}}\ldots\text{Object}_{\text{Embedded}}\ \text{Verb}_{\text{Embedded}}\ldots\text{Object}_{\text{Main}}\ \text{Verb}_{\text{Main}}$. A degree-0 learner would need to track information spanning the embedded clause. A learner with the cognitive resources to do that would most likely also have the cognitive resources to track the information in the embedded clause. So, a degree-0 learner that is motivated by a limit on cognitive resources and who must learn a head-final language might be redefined as one using the information in the portion of the degree-0 clause that is adjacent, i.e. not split by any embedded clause material.

[26] Populations canonically grow at an exponential rate, with the current set of new population members typically outnumbering the previous set of new population members. The exact amount that the current set of new members outnumbers the previous set of new members is described by the population growth coefficient, a constant value specific to a given population.

probability distribution for the language as a whole. The V2 rule’s restriction to matrix clauses means that while the distribution of clauses in the matrix is mixed between VO and OV order, Old English (before the change) is strongly OV in embedded clauses (see table 4.1 in section 4.4.1.2). This is a case where unambiguous data and degree-0 data filters on data intake should create a mismatch between the adult’s underlying probability distribution and the probability distribution the child converges on.

Since we have historical records allowing us to calculate the rate of change from OV to VO, I model the effect of filtering by restricting my model to learn from simple unambiguous structures in the quantities found in the historical record at the beginning of the transformation of Old English from OV to VO. The model will then create a set of successive generations, each diverging from the initial distribution to a designated extent; this is the rate of change. Then, I can calculate the effect of these two filters on the rate of change in the model, and compare it to the actual rate calculated from the distribution of data found at various periods during this transformation in the actual historical record.

I do this in two steps. First, I ask if a population whose learners filter their input down to degree-0 unambiguous data is able to follow the historically attested trajectory. Then I ask whether a model that uses additional data (ambiguous or embedded or both) during learning could also produce the observed historical patterns in the simulated population. This provides us with the evidence we need to determine if children should use these filters during language learning.

4.3 Old English

4.3.1 OV and VO word order in Old English

Between 1000 A.D. and 1150 A.D., the distribution in the Old English population consisted of mostly OV order utterances (11a) while the distribution in the population at 1200 A.D. consisted of mostly VO order utterances (11b) (YCOE Corpus, Taylor et al., 2003; PPCME2 Corpus, Kroch & Taylor, 2000).

(11a)	he$_{\text{Subj}}$	Gode$_{\text{Obj}}$	þancode$_{\text{TensedVerb}}$
	he	*God*	*thanked*

‘He thanked God’
(*Beowulf,* 625, ~1100 A.D.)

(11b)	& [mid his stefne]$_{\text{PP}}$	he$_{\text{Subj}}$	aweců$_{\text{TensedVerb}}$	deade$_{\text{Obj}}$	[to life]$_{\text{PP}}$
	& with his stem	*he*	*awakened*	*the-dead*	*to life*

“And with his stem, he awakened the dead to life.”
(*James the Greater,* 30.31, ~1150 A.D.)

4.3.2 Unambiguous Data

4.3.2.1 Unambiguous OV

Unambiguous data for OV word order correlate with observable data of the following types in Old English: (12a) the tensed Verb appears at the end of the clause or (12b) the non-tensed Verb remains in the post-Object position, while the tensed auxiliary moves.

(12a) he$_{\text{Subj}}$ hyne$_{\text{Obj}}$ gebidde$_{\text{TensedVerb}}$
He him may-pray
'He may pray (to) him'
(*Ælfric's Letter to Wulfsige,* 87.107, ~1075 A.D.)

(12b) we$_{\text{Subj}}$ sculen$_{\text{TensedVerb}}$ [ure yfele þeawes]$_{\text{Obj}}$ forlæten$_{\text{Non-TensedVerb}}$
we should our evil practices abandon
'We should abandon our evil practices.'
(*Alcuin's De Virtutibus et Vitiis,* 70.52, ~1150 A.D.)

4.3.2.2 Unambiguous VO

A reasonable assumption might be that unambiguous VO data should be the counterpart of unambiguous OV data in form. Specifically, one might assume that since *Subject Object TensedVerb* is unambiguous OV data, *Subject TensedVerb Object* should then be unambiguous VO data. However, recall the V2 movement rule, which moves the tensed Verb to the second phrasal position of the clause. As we will see below, when this movement rule is taken into account, sentences of the form *Subject TensedVerb Object* cannot be perceived as unambiguous VO data.

4.3.2.2.1 V2 Interference

Assuming V2, a simple *Subject TensedVerb Object* utterance could be parsed with either the OV (with V2 movement) or the VO order parameter value (with or without V2 movement). Example (13) shows this: the tensed Verb *clænsað* could begin in sentence final position (OV order) and move to the second position (13a), or it could be generated in this position all along (VO order) (13b).

(13a) heo$_{\text{Subj}}$ clænsað$_{\text{TensedVerb}}$ t_{Subj} [þa sawle þæs rædendan]$_{\text{Obj}}$ $t_{TensedVerb}$
they purified the souls [the advising]-Gen

(13b) heo$_{\text{Subj}}$ clænsað$_{\text{TensedVerb}}$ [þa sawle þæs rædendan]$_{\text{Obj}}$
they purified the souls [the-advising]-Gen
'They purified the souls of the advising ones.'
(*Alcuin's De Virtutibus et Vitiis*, 83.59, ~1150 A.D.)

Because of V2 movement, unambiguous VO data in matrix clauses appears as the examples in (14): there is either (a) more than one phrase to the left of the Verb ([*mid his stefne*]$_{\text{PP}}$ *he*$_{\text{Subj}}$), ruling out a V2 analysis, or (b) some sub-piece of the verbal complex (*up*$_{\text{Verb-Marker}}$) immediately preceding the Object.

(14a) & [mid his stefne]$_{\text{PP}}$ he$_{\text{Subj}}$ aweceð$_{\text{TensedVerb}}$ deade$_{\text{Obj}}$ [to life]$_{\text{PP}}$
& with his stem he awakened the-dead to life
'And with his stem, he awakened the dead to life.'
(*James the Greater,* 30.31, ~1150 A.D.)

(14b) þa$_{\textbf{Adv}}$ ahof$_{\text{TensedVerb}}$ Paulus$_{\text{Subj}}$ up$_{\text{Verb-Marker}}$ [his heafod]$_{\text{Obj}}$
then lifted Paul up his head
'Then Paul lifted his head up.'
(*Blickling Homilies,* 187.35, between 900 and 1000 A.D.)

4.3.2.2.2 Verb-Markers

I will term sub-pieces of the verbal complex "Verb-Markers". A Verb-Marker is a word that is semantically associated with a Verb, such as a particle ('up', 'out'), a non-tensed complement to tensed Verbs, a closed-class adverbial ('never'), or a negative ('not') (Lightfoot, 1991). Under the assumption that the learner believes all Verb-like words should be adjacent to each other (Lightfoot, 1991), a Verb-Marker can be used to determine the original position of the Verb. For (14b), the Verb-Marker *up* indicates the position where the tensed Verb originated before V2 movement; since the Verb-Marker precedes the Object, the original position of the Verb is assumed to be in front of the Object as well. So, this utterance type is perceived as unambiguous data for VO order. Examples of utterances with Verb-Markers are in (15) below (Verb-Markers are in bold): the particle *up* is a Verb-Marker in (15a) and the non-tensed Verb *gewyrecean* is a Verb-Marker in (15b).

(15a) þa$_{\textbf{Adv}}$ ahof$_{\textbf{TensedVerb}}$ Paulus$_{\textbf{Subj}}$ **up**$_{\textbf{Particle}}$ [his heafod]$_{\text{Obj}}$
then lifted Paul ***up*** *his head*
'Then Paul lifted his head up.'
(*Blickling Homilies,* 187.35, between 900 and 1000 A.D.)

(15b) Swa$_{\textbf{Adv}}$ sceal$_{\textbf{TensedVerb}}$ [geong guma]$_{\textbf{Subj}}$ gode$_{\textbf{Obj}}$
Thus shall young men good-things
gewyrecean$_{\textbf{Non-TensedVerb}}$
perform
'Thus shall young men perform good things.'
(*Beowulf,* 20, ~1100 A.D.)

Interestingly, Old English Verb-Markers (unlike their modern Dutch and German counterparts) were *unreliable* as a marker of the Verb's original position. In many cases (such as the negative *ne* in (15c) below), the Verb-Marker would not remain adjacent to the Object. If there were no other Verb-Markers adjacent to the Object, then no indication of the Verb's initial position remained and the utterance could be interpreted as ambiguous between the OV or VO order hypotheses. In (15c), the adverbial *næfre* remains adjacent to the Object, and so this data point would be perceived as unambiguous for VO order.

(15c)	**ne**$_{\textbf{Negative}}$	geseah$_{\text{TensedVerb}}$	ic$_{\text{Subj}}$	**næfre**$_{\textbf{Adverbial}}$	[ða burh]$_{\text{Obj}}$
	NEG	*saw*	*I*	***never***	*the city*

'Never did I see the city.'
(Ælfric, *Homilies.* I.572.3, between 900 and 1000 A.D.)

4.3.3 Causes of Language Change

4.3.3.1 The Effect of the Unambiguous Data Distribution

As we have just seen, matrix clause cues (such as the location of a Verb-Marker with respect to the Object) can be unreliable. This causes data that potentially could have been perceived as unambiguous to be perceived as ambiguous. Thus, a learner using an unambiguous data filter would potentially encounter a distribution of OV and VO data points that is different from the distribution the adult speakers of the population used to generate the entire data set. In short, the learner's intake can have a different distribution than that of the available input. This difference in the intake can cause successive generations of Old English children to have different OV/VO probability distributions than their predecessors. The Old English population would then shift to a strongly VO-biased distribution because of what the learners' intake consists of. I will formally model this intuition by using actual quantitative data from the relevant historical periods coupled with an explicit probabilistic model.

4.3.3.2 A Concern About Other Causes of Language Change

Before we examine the details of the model, I should address a concern about the cause of this particular language change in Old English. I have assumed, based on Lightfoot's (1991) claim, that language learning (an internal factor) is the instigator of the shift from a strongly OV-biased distribution to a strongly VO-biased distribution. However, one might wonder if external factors could have played a more significant role in this change.

I consider two potential external factors below: Scandinavian influence and Norman influence. We will see that neither factor *by itself* could have caused the change in Old English from a strongly OV-biased distribution to a strongly VO-biased distribution. However, it is still possible that the correct combination and influence of external factors could have produced the recorded historical change, even in the absence of the imperfect learning approach advocated by Lightfoot (1991) and adopted here. The contribution of the present work would then be to demonstrate how it is not *necessary* to have external factors in order to cause abrupt change at the population-level in such a limited timeframe.

4.3.3.2.1 Scandinavian Influence

Scandinavian influence before 1000 A.D. is claimed to have caused Old English Verb-Markers to become unreliable (Kroch & Taylor, 1997). Old Norse, the language spoken by the Scandinavians, used VO order and therefore introduced variability into the OV ordered Old English. Is it possible that continued Scandinavian influence *alone* caused the sharp change in the OV/VO distribution of Old English between 1150 A.D. and 1200 A.D.? To accomplish this, a continuous stream of Scandinavian speakers would be the force that caused the overall composition of the Old English population to drift towards a VO-biased distribution by 1150 A.D. These Scandinavians would learn Old English as a second language, and therefore likely learn it imperfectly, perhaps introducing a continuous VO bias into the data set available to learners in the population.

Old English learners, not filtering the input, would simply converge on exactly the distribution they encountered in the input from the mixture of native Old English and Scandinavian speakers using Old English as a second-language. This scenario, however, would require an exponential increase of incoming Scandinavians in order to get the gradual population-level shift before 1150 A.D. and the sharp population-level shift after 1150 A.D. This seems to be a rather unlikely event.

Still, there is another variant on this external factor. Suppose there was some prestige associated with the Scandinavians such that Old English speakers altered their OV/VO usage to accommodate (see Giles & Powesland (1975) for accommodation theory) and sound more like the Scandinavian portion of the population. So, Scandinavians would be learning Old English as a second language from native Old English speakers who would be more VO-biased (as a conscious social effort). The overall composition of the population would then be increasingly more VO-biased as time went on. Yet, in order to achieve the historical S-shaped trajectory of change, again there needs to be an exponential increase somewhere – either in the number of Scandinavians joining the Old English population or in the associated prestige with the Scandinavian VO-bias. While less unlikely than the previous scenario, relying on an exponential increase of Scandinavian prestige doesn't seem ideal as the sole factor driving change, either.

Nonetheless, we should not discount Scandinavian influence completely. Scandinavian influence combined with input filtering could well give the desired change. Later in this chapter, we will see that adult utterances generated with OV order are more prone than their VO counterparts to becoming ambiguous in the observable data. Scandinavian influence, being VO-biased, could have been responsible for this. Thus, learners using an unambiguous data filter would have become more VO-biased over time since the VO data generated by the Old English speakers was less likely to become ambiguous. Still, it is crucial to note that this scenario is the result of the combination of Scandinavian influence and language change caused by language learning. Scandinavian influence alone seems unlikely to be the cause of the language change in Old English.

4.3.3.2.2 Norman Influence

A second external source of influence is the Norman invasion in 1066 A.D. The Norman invaders spoke Old French, which was OV-biased in its distribution (Kibler 1984): embedded clauses were predominantly OV order, as well as the matrix clauses. So, contact with Old French speakers would have biased the Old English population to become more OV. However, between 1000 and 1150 A.D., the Old English population was already drifting towards being more VO in its distribution. So, any contact with Old French speakers would have hindered the population-level change to a VO-biased distribution. This influence may have been tempered (and overcome) by the VO-biased Scandinavian influence.

Another possibility is disaccommodation with the OV-biased distribution from the Old French speakers if there was social stigma associated with the language of the Norman invaders (again, see Giles & Powesland (1975) for accommodation theory). Old English speakers, disliking the invaders (and perhaps liking the Scandinavians) would be driven to more VO-biased usage. Still, it remains clear that contact with the Normans alone could not have *caused* the shift in Old English to a strongly VO-biased distribution unless, as discussed for the Scandinavian influence in the previous section, there was an exponential increase somewhere – in this case, in the social stigma associated with using an OV-biased distribution.

4.4 The Model

I now describe the model at the individual level and the population level. Because I have posited that language change at the population level is driven by language learning at the individual level, I first examine the details of individual learning. In the model, the learner has different hypotheses for a structure in the language (such as OV and VO word order) available during learning, in line with work by Yang (2002), Dresher (1999), Lightfoot (1999), Fodor (1998a, 1998b), Niyogi & Berwick (1997, 1996, 1995), and Clark & Roberts (1993). The target state after learning is complete is a probabilistic distribution between competing hypotheses (Yang, 2002; Pintzuk, 2002; Kroch & Taylor, 1997; Bock & Kroch, 1989). Because of this, individual linguistic behavior, whether child (Yang, 2003) or adult (Bock & Kroch, 1989), is represented as a probabilistic distribution of multiple structural hypotheses, specifically between OV and VO word order.

Population-level change in the model is the result of a build-up of individual-level "mislearning" (Yang, 2002, 2000; Briscoe, 2000, 1999; Niyogi & Berwick, 1997, 1996, 1995; Clark & Roberts, 1993; Lightfoot, 1991). Thus, the population-level model relies heavily upon the individual-level implementation.

4.4.1 The Individual-Level Model

4.4.1.1 Learning in the Individual

The individual-level model is a model of language learning. An individual learner in the model is instantiated with a probability p_{VO} of accessing VO word order. The OV word order is accessed with probability 1 – p_{VO}, as there are only two hypotheses under consideration.

In a language system where the adult speakers have $p_{VO} = 1.0$ (modern English) or $p_{VO} = 0.0$ (modern Dutch and German), all utterances are produced with one word order (VO for modern English, OV for modern Dutch and German). This directly impacts the distribution of unambiguous data, since all unambiguous data will be unambiguous for a single hypothesis (either OV words order or VO word order).

In contrast, a language system can also exist where the adult p_{VO} is greater than 0.0 and less than 1.0, such as the state of Old English between 1000 A.D. and 1200 A.D. In a system like Old English, VO order is accessed for production with probability p_{VO} (which is less than 1.0) and the OV order is accessed with probability 1-p_{VO} (which is greater than 0.0). This will impact the distribution of unambiguous data: the data will have some distribution between $p_{VO} = 0.0$ (all OV order data) and $p_{VO} = 1.0$ (all VO order data). The learner then determines her own p_{VO} based on the distribution in the intake (which, in the model, will be filtered down to the degree-0 unambiguous data).

The model assumes no initial bias for either hypothesis, so the initial value for a learner's word order, p_{VO}, is 0.5. This can be interpreted as an unbiased value, since it is precisely in the middle of $p_{VO} = 0.0$ (all OV order) and $p_{VO} = 1.0$ (all VO order). Note that an unbiased p_{vo} would predict that very young children of any language would have an unstable word order initially. I speculate that the reason why children always demonstrate knowledge of the correct word order by the time they reach the two word stage is because they have already been exposed to enough examples of the appropriate word order for their language to bias them in the correct way.

The final p_{VO} value after the learning period is complete will range between 0.0 and 1.0, and can be interpreted as a probabilistic access of the OV and VO words orders. A p_{VO} of 0.3, for example, would correspond to accessing VO order 30% of the time during production and OV order 70% of the time.

Since the initial p_{VO} for the learner is 0.5, the learner initially expects the distribution of OV and VO data in the intake to be unbiased. I use the Bayesian framework laid out in chapter 2 to model how the learner's initial hypothesis about the OV/VO distribution (p_{VO}) shifts with each additional data point from the intake. In addition to the support for its psychological validity in human cognition (Tenenbaum & Griffiths, 2001), Bayesian learning has also been used in other models of language evolution and change (Briscoe, 1999).

Since there are only two values for the OV/VO ordering (OV and VO), I represent the learner's hypothesis of the expected distribution of OV and VO utterances as a binomial distribution centered around some probability *p*. Here, probability *p* is p_{VO} and represents the learner's belief about the likelihood of encountering a VO utterance. When p_{VO} is 0.5, the learner is most confident that it is equally likely that an OV or the VO utterance will be encountered. A p_{VO} near 0.0 means the learner is most confident that a VO utterance will never be encountered; a p_{VO} near 1.0 means the learner is most confident that a VO utterance will always be encountered.

The learner's p_{VO} is updated by calculating the maximum of the a posteriori (MAP) probability of the prior belief p_{VOprev}, given the current piece of data from the intake. In essence, the model is starting with a prior probability and its expected distribution of OV and VO utterances, and comparing this expected distribution against the actual distribution encountered. The updated probability is calculated as follows:

(16a) If the data point is analyzed as OV, $p_{VO} = (p_{VOprev}*t)/(t+c)$
(16b) If the data point is analyzed as VO, $p_{VO} = (p_{VOprev}*t+c)/(t+c)$

where t = total expected number of data points in the *intake* during the period of fluctuation (2000 in this model) and c = learner's confidence in the input (ranging between 0.0 and 5.0), based on p_{VOprev}. Note that t refers to quantity of data points in the intake, and not the input. Thus, the learner will encounter considerably more than 2000 data points in the input; the fluctuation period, however, ends when 2000 data points from the intake have been encountered.

Also note that these equations are a modification of the update equations derived in chapter 2. In those equations, $c = 1$. However, I have modified this value since those equations (with $c = 1$) would not allow the learner to converge to 1.0 or 0.0, even if all unambiguous data are of one value. For example, with $t = 2000$, encountering all OV data points causes the final p_{VO} to be 0.194 (not 0.0); encountering all VO data points causes the final p_{VO} to be 0.816 (not 1.0). I therefore modified c to allow the final p_{VO} to be closer to the endpoint values (either 0.0 or 1.0) for each case.

The value c ranges linearly between 0 and a maximum value m, depending on what p_{VOprev} is[27]:

(17a) VO data: $c = p_{VOprev} * m$
(17b) OV data: $c = (1 - p_{VOprev}) *m$

The value m ranges between 3.0 and 5.0. The m for a particular mixture of degree-0 and degree-1 data is determined by seeing which m value allows the simulated Old English population to reach an average p_{VO} value in the population between 1000 and 1150 A.D that accords with the historical data available. For example, the value of m for an intake that consists only of degree-0 data is 5.0.

With the new update functions, unambiguous data for one value the entire time will cause the final p_{VO} to be much closer to the endpoint. Seeing 2000 OV data points leaves p_{VO} between .007 and .048 (depending on m); seeing 2000 VO data points leaves p_{VO} between .952 and .993 (depending on m).

The final p_{VO} at the end of the fluctuation period (after t data points from the intake have been encountered) will reflect the distribution of the data points in the intake. Importantly, the distribution is reflected *without* the learner explicitly memorizing each individual piece of data for later analysis. Instead, as each data

[27] The same effect could likely be achieved by holding c between 0 and 1, and letting t vary. However, this loses the intuition that t (the number of data points the learner expects, i.e. the amount of change allowed) should be the same across the different conditions investigated.

point is encountered, the information is extracted from that data point and, using the equations in (16) and (17), integrated into the learner's hypothesis about what the distribution of OV and VO data points is expected to be.

The individual learning algorithm used in the model is described in (18):

(18) Individual learning algorithm

(a) Set initial p_{VO} to 0.5.

(b) Get a data point from an "average" member of the population. The input for the learner is determined by sampling from a normal distribution around the average p_{VO} of the population.

(c) If the data point is degree-0 and unambiguous, use this data point as intake and then alter p_{VO} accordingly.

(d) Repeat (b-c) until the fluctuation period is over, as determined by *t*.

For each data point encountered from the input, the learner determines if the data point belongs in the intake. If so, p_{VO} is updated using the equations in (16-17). This process of encountering input and integrating the information from data in the intake continues until the fluctuation period is over. At that point, the learner becomes one of the population members that contribute to the average p_{VO} value that will influence future learners. The higher the average p_{VO} value is in the population, the more likely learners are to encounter unambiguous VO data.

4.4.1.2 Old English Intake Data

As we have seen, the distribution in the learner's intake controls the learner's shift away from the unbiased probability of $p_{VO} = 0.5$. The only way to shift p_{VO} away from 0.5 is to have more data points of one word order than of the other in the intake. I will refer to this quantity as the bias one word order has over the other.[28] So, if the intake distribution is OV-biased, there are more OV data points in the learner's intake. If the intake distribution is VO-biased, there are more VO data points in the learner's intake. Note that if the intake is a subset of the input (due to filtering), the bias with respect to the available input is smaller than the bias with respect to the learner's intake. Table 4.1 displays the OV bias with respect to the input in the degree-0 (D0) and degree-1 (D1) clauses in Old English at various points in time.

[28] This differs from the *advantage* (Yang, 2000) one hypothesis has over another. Advantage there is defined as inherent grammar incompatibility – one hypothesis will have an advantage when the opposing hypothesis is incompatible with data *types*. Thus, it does not matter for advantage how frequent a data type is, e.g. how many data *tokens* appear in the intake. It simply matters that there are data types one hypothesis is incompatible with. Advantage is thus different from the *bias* in the intake distribution, which very much depends on the quantity of data tokens that are unambiguous for one hypothesis vs. the other. More specifically, a hypothesis with a lower advantage can still have a stronger bias in the data intake distribution, and vice versa.

	D0 Total # Clauses	D0 Unamb OV	D0 Unamb VO	D0 OV Bias w.r.t. the input[a]	D0 OV Bias w.r.t. the intake[b]
1000 A.D.	9805	1389	936	4.6%	19.5%
1000 – 1150 A.D	6214	624	590	0.5%	2.8%
1200 A.D.	1282	180	190	-0.8% [c]	-2.7% [c]

	D1 Total # Clauses	D1 Unamb OV	D1 Unamb VO	D1 OV Bias w.r.t. the input[a]	D1 OV Bias w.r.t. the intake[b]
1000 A.D.	7559	3844	1583	29.9%	41.7%
1000 – 1150 A.D	3636	1759	975	21.6%	28.7%
1200 A.D.	2236	551	1460	-40.7% [c]	-45.2% [c]

Table 4.1. OV order bias in the input for degree-0 (D0) and degree-1 (D1) clauses.
[a] The bias for the OV order with respect to the *input* is derived by subtracting the quantity of VO data from the quantity of OV data, and then dividing by the total number of clauses in the input. For instance, the D0 OV bias at 1000 A.D. is calculated as (1389-936)/9805 = 4.6%. [b] The bias for the OV order with respect to the *intake* is derived by subtracting the quantity of unambiguous VO data from the quantity of unambiguous OV data, and then dividing by the total number of clauses in the intake. For instance, the D0 OV bias at 1000 A.D. is calculated as (1389-936)/(1389+936) = 19.5%. [c] Note that a negative OV bias means that the distribution is VO-biased.

The corpus data show a 4.6% bias with respect to the input for the OV order in the degreee-0 clauses at 1000 A.D. We can interpret this as less than 5 out of every 100 sentences of the available input are biasing the learner away from a p_{VO} of 0.5 (and towards an OV value of 0.0). With respect to the intake, the OV order bias is much higher: just about 1 out of every 5 data points in the intake biases the learner towards 0.0 (OV order).

Interestingly, the OV bias in the degree-1 clauses is much higher (29.9% with respect to the input, and 41.7% with respect to the intake). However, a degree-0 filter would cause the learner to ignore these data that would shift p_{VO} towards 0.0 significantly more often. Nonetheless, the difference of the bias in the different distributions highlights the effect that data intake filtering can have: the bias in the distribution alters quite a lot depending on which data set the learner is using.

4.4.2 Population-Level Model for Old English

4.4.2.1 Population-Level Algorithm and Population Growth

The population algorithm (19) centers on the individual acquisition algorithm in (18).

(19) Population-level algorithm

(a) Set the age range of the population from 0 to 60 years old and create 18,000 population members.
(b) Initialize the members of the population to the average p_{VO} at 1000 A.D.
(c) Set the time to 1000 A.D.
(d) Move forward 2 years.
(e) Members age 59-60 die off. The rest of the population ages 2 years.
(f) New members are born. These new members use the individual acquisition algorithm (18) to set their p_{VO}.
(g) Repeat steps (d-f) until the year 1200 A.D.

The population members range in age from newborn to 60 years old.[29] The initial size of the population is 18,000, based on estimates from Koenigsberger & Briggs (1987). At 1000 A.D., all the members of the population have their p_{VO} set to the same initial p_{VO}, which is derived from the historical corpus data. Every two years, new members are born to replace the members that died as well as to increase the overall size of the population so it matches the growth rate extrapolated from Koenigsberger & Briggs (1987). Populations are estimated to grow at an exponential rate characterized by the equation in (20).

(20) Population growth equation

population size = previous population size * e^{rt}

For the Old English population in our model, $r = 0.00400953$ and t = time in years. For example, at 1002 A.D., the estimated population size is $18000 * e^{0.00400953*2} = 18145$. Thus, once the oldest members (age 59-60) die off, enough new members are born to make the total population size at 1002 A.D. be 18145. These new members encounter input from the rest of the population and follow the process of individual acquisition laid out previously in order to determine their final p_{VO}. This process of death, birth, and learning continues until the year 1200 A.D.

4.4.2.2 Population Values from Historical Data

I use the historical corpus data to initialize the average p_{VO} in the population at 1000 A.D., calibrate the model between 1000 and 1150 A.D. (recall that the confidence value c in update equation (16) needs calibration), and determine how strongly VO-biased the distribution has to be in the population by 1200 A.D. But it is not straightforward to determine the average p_{VO} at a given period of time.

[29] The population members begin uniformly distributed between 0 and 60 years old, though this could easily be modified to a more skewed distribution where there are more younger members of the population than older. In addition, the age maximum (60 years old) was arbitrarily chosen. Having a lower maximum (say, 40 years old) would possibly speed the rate of change through the population. However, the overall results would likely be the same as found here since the population model must be calibrated so that the population remains sufficiently OV-biased before 1150 A.D. That is, a sufficient OV-bias in the population before 1150 A.D. is a precondition. The behavior we are interested in is how a population that is sufficiently OV-biased before 1150 A.D. changes between 1150 A.D. and 1200 A.D. Specifically, can it become VO-biased enough?

Both the degree-0 and degree-1 unambiguous data distributions are likely to be distorted from the underlying unambiguous data distribution produced by p_{VO} because the degree-0 and degree-1 clauses have ambiguous data. The underlying distribution in a speaker's mind, however, has no ambiguous data – every clause is generated with OV or VO order. As we can see in table 4.2, the degree-0 clauses have more ambiguous data than the degree-1 clauses. Moreover, recall from table 1 that the degree-1 clauses also have a magnified bias, compared to the degree-0 clauses. Taken together, I use these two observations to make the assumption that the degree-0 distribution is more distorted than the degree-1 distribution.

	D0 Total # Clauses	D0 # Unamb Clauses	D0 % Ambig[a]
1000 A.D.	9805	2325	(9805-2325)/9805 = 76%
1000-1150 A.D.	6214	1214	(6214-1214)/6214 = 80%
1200 A.D.	1282	370	(1282-370)/1282 = 71%

	D1 Total # Clauses	D1 # Unamb Clauses	D1 % Ambig[a]
1000 A.D.	7559	5427	(7759-5427)/7759 = 28%
1000-1150 A.D.	3636	2734	(3636-2734)/3636 = 25%
1200 A.D.	2236	2011	(2236-2011)/2236 = 10%

Table 4.2. Percentage of ambiguous clauses in the historical corpora. [a] The % Ambig is calculated by dividing the number of ambiguous clauses (Total - Unamb) by the total number of clauses.

I then use the difference in distortion between the degree-0 and degree-1 unambiguous data distributions to estimate the difference in distortion between the degree-1 distribution and the underlying unambiguous data distribution in a speaker's mind. In this way, I estimate the underlying unambiguous data distribution (produced by p_{VO}) for an average Old English speaker at certain points in time.

I will first step through the formalization of the procedure used to derive the underlying p_{VO} at a given point in time. Then, I will step through an explicit example from the Old English historical data.

4.4.2.2.1 Procedure to Derive p_{VO} from Historical Data

Let there be two hypotheses under consideration, h1 and h2. For Old English, these are OV order (h1) and VO order (h2). From historical corpora, we can gather unambiguous data points for h1 and h2 in both the degree-0 and degree-1 clauses. From these, we can calculate the number of ambiguous data points in the degree-0 and degree-1 clauses. The quantities gathered from historical corpora are **u1d0** (**u**nambiguous data points for h**1** in **d**egree-**0** clauses), **u2d0** (**u**nambiguous data points for h**2** in **d**egree-**0** clauses), **ad0** (**a**mbiguous data points in **d**egree-**0** clauses), **u1d1** (**u**nambiguous data points for h**1** in **d**egree-**1** clauses), **u2d1** (**u**nambiguous data points for h**2** in **d**egree-**1** clauses), and **ad1** (**a**mbiguous data points in **d**egree-**1** clauses) in table 4.3 below. The quantities that must be derived are *u1* and *u2*, which represent the quantities of unambiguous data for each hypothesis in the underlying distribution that the average population speaker produced. In the underlying distribution, there are no ambiguous data because the speaker either accesses h1 or h2 to produce the data point. Once $u1$ and $u2$ are known, p_{VO} can be derived ($p_{VO} = u2/(u1 + u2)$).

	Unamb OV (h1)	Unamb VO (h2)	Amb
Degree-0	**u1d0**	**u2d0**	**ad0**
Degree-1	**u1d1**	**u2d1**	**ad1**
Underlying Distribution	*u1*	*u2*	0

Table 4.3. Formalization of quantities available from historical corpora and quantities to derive. Quantities in **bold** can be gathered from historical corpora. Quantities in *italics* must be derived and are used to calculate the average p_{VO} in the population.

Let γ represent the probability that the speaker accesses h1 during production. Since there are only two options under consideration, $1 - \gamma$ represents the probability the speaker accesses h2 during production.

Let the total quantity of degree-0 data be **d0**. So, **d0** = **u1d0 + u2d0 + ad0**.

Let the total quantity of degree-1 data be **d1**. So, **d1** = **u1d1 + u2d1 + ad1**.

We first must normalize the degree-1 data quantity to the degree-0 data quantity. After normalization, **u1d1' + u2d1' + ad1' = d0 = u1d0 + u2d0 + ad0**.

(21) Equation quantities, original and normalized

(a) **d0** = **u1d0 + u2d0 + ad0**

(b) **d1** = **u1d1 + u2d1 + ad1**

(c) **d0 = u1d1' + u2d1' + ad1'**

	Unamb OV (h1)	Unamb VO (h2)	Amb
Degree-0	**u1d0**	**u2d0**	**ad0**
Degree-1	**u1d1'** **= u1d1*(d0/d1)**	**u2d1'** **= u2d1*(d0/d1)**	**ad1'** **= ad1*(d0/d1)**
Underlying Distribution	*u1*	*u2*	0

Table 4.4. Data quantities after normalization.

The value $u1$ represents the quantity of unambiguous h1 (OV) data generated by the speaker. The value $u2$ represents the quantity of unambiguous h2 (VO) data generated by the speaker. Since there are no ambiguous data, let these two quantities also sum to **d0** ($u1 + u2 =$ **d0**). This represents the intuition that $u1$ and $u2$ have been "normalized" so that they can be compared against their counterpart values in the degree-1 and degree-0 distributions. Note that since $u1$ and $u2$ have not been calculated yet, we can simply make them sum to the appropriate normalized value, **d0**.

(22) Underlying distribution “normalization”

$u1 + u2 = \mathbf{d0}$

Recall that the probability that a speaker accesses h1 when producing a data point is γ. Since the total quantity of unambiguous data points in the underlying distribution has been normalized to **d0**, this probability can now be set equal to $u1/\mathbf{d0}$. Thus, we can rewrite *u1* as $\gamma*\mathbf{d0}$.

(23) Rewriting underlying distribution quantity *u1*

$\gamma = u1/\mathbf{d0}$
$u1 = \gamma*\mathbf{d0}$

I now make an assumption about the relation of underlying data distribution to the degree-1 data distribution. Specifically, I assume that the degree-1 distribution originally had the same number of h1 data points as the underlying distribution, but that some of these data points became ambiguous (due to various grammatical operations). Thus, we can relate the underlying distribution h1 data point quantity *u1* to the degree-1 data quantities **u1d1’** (normalized quantity of **u**nambiguous data points for h**1** in the **d**egree-**1** distribution) and **ad1’** (normalized quantity of **a**mbiguous data points in the **d**egree-**1** distribution).

(24) Relation between *u1* and **u1d1’** and **ad1’**

$u1 = \mathbf{u1d1'}$ + the portion of **ad1’** that were originally h1 data points
Let $a1d1$ = portion of **ad1’** that were originally h1 data points
$u1 = \mathbf{u1d1'} + a1d1$

Recall from (23) that *u1* can be rewritten in terms of γ and **d0**. We can thus write an equation for *a1d1*, the portion of **ad1’** that were originally h1 data points.

(25) Writing an equation for *a1d1*

$u1 = \gamma*\mathbf{d0}$ (from (23))
$\gamma*\mathbf{d0} = \mathbf{u1d1'} + a1d1$ (from (24))
$a1d1 = \gamma*\mathbf{d0} - \mathbf{u1d1'}$

Since there are only two hypotheses, the portion of **ad1’** that were not originally h1 data points must have been h2 data points. Given this, we can write an equation for *a2d1*, the portion of **ad1’** that were originally h2 data points.

(26) Writing an equation for *a2d1*

Let $a2d1$ = portion of **ad1’** that were originally h2 data points
$\mathbf{ad1'} = a1d1 + a2d1$
$a2d1 = \mathbf{ad1'} - a1d1$

Moreover, using the same assumption as before about the relation between the underlying distribution and the degree-1 distribution, we can rewrite *u2*, the quantity of unambiguous data points for h2 in the underlying distribution.

(27) Rewriting *u2*

u2 = **u2d1’** + the portion of **ad1’** that were originally h2 data points
a2d1 = portion of **ad1’** that were originally h2 data points
u2 = **u2d1’** + *a2d1*
u2 = **u2d1’** + **ad1’** – *a1d1*   (from (26))

	Unamb OV (h1)	Unamb VO (h2)	Amb
Degree-0	**u1d0**	**u2d0**	**ad0**
Degree-1	**u1d1’** **= u1d1*(d0/d1)**	**u2d1’** **= u2d1*(d0/d1)**	**ad1’ =** **ad1*(d0/d1)**
Underlying Distribution	**u1d1’** + *a1d1*	**u2d1’** + (**ad1’** – *a1d1*)	0

Table 4.5**.** Derived quantities rewritten.

Now, we look at the relation between the degree-1 and the degree-0 distribution. I make an assumption similar to the one we did about the relation between the underlying distribution and the degree-1 distribution: specifically, I assume that the degree-0 distribution originally had the same number of h1 or h2 data points as the degree-1 distribution, but that some of these data points became ambiguous (due to various grammatical operations). I can describe these quantities in terms of values we have already observed or calculated.

I assume that the quantity of h1 data points in the degree-0 distribution was originally the same as the quantity of h1 data points in the normalized degree-0 distribution, **u1d1’**. However, some became ambiguous and only **u1d0** remain. So, the quantity of data points that became ambiguous going from the degree-1 distribution to the degree-0 distribution can be described as **u1d1’** – **u1d0**. The same reasoning can be used for the h2 data points.

(28) Quantities of data points that became ambiguous going from the degree-1 distribution to the degree-0 distribution

Let the quantity of h1 data points that became ambiguous going from the degree-1 to the degree-0 distribution = *a1d1to0.*

a1d1to0 = **u1d1’ – u1d0**

Let the quantity of h2 data points that became ambiguous going from the degree-1 to the degree-0 distribution = *a2d1to0*

a2d1to0 = **u2d1’ – u2d0**

We can now define an *ambiguity loss ratio* **Ld1to0**, which represents the ratio of h1 data points that became ambiguous compared to the h2 data points that became unambiguous going from the degree-1 to the degree-0 distribution.

(29) Ambiguity Loss Ratio **Ld1to0**
(h1 data point loss over h2 data point loss going from degree-1 to degree-0 distribution)

$$\mathbf{Ld1to0} = \frac{\mathbf{u1d1'\ -\ u1d0}}{\mathbf{u2d1'\ -\ u2d0}}$$

We can then describe the quantities of h1 and h2 data points that become ambiguous going from the underlying distribution to the degree-1 distribution. Let *a1utod1* be the quantity of h1 data points that became ambiguous going from the underlying distribution to the degree-1 distribution. Let *a2utod1* be the quantity of h2 data points that become ambiguous going from the underlying distribution to the degree-1 distribution.

(30) Describing the quantities of h1 and h2 data points that become ambiguous going from the underlying to the degree-1 distribution

(a) *a1utod1* (h1 data points that become ambiguous)

$a1utod1 = u1 - \mathbf{u1d1'}$

$a1utod1 = (\mathbf{u1d1'} + a1d1) - \mathbf{u1d1'}$ (from (24))

$a1utod1 = a1d1$

(b) *a2utod1* (h2 data points that become ambiguous)

$a2utod1 = u2 - \mathbf{u2d1'}$

$a2utod1 = (\mathbf{u2d1'} + (\mathbf{ad1'} - a1d1)) - \mathbf{u2d1'}$ (from (27))

$a2utod1 = \mathbf{ad1'} - a1d1$

We can now define an ambiguity loss ratio **Lutod1**, which represents the ratio of h1 to h2 data points that become "lost" to ambiguity going from the underlying distribution to the degree-1 distribution. I make an assumption that **Ld1to0** is the same as **Lutod1**, that is that the rate at which h1 data points become ambiguous compared to h2 data points does not change depending on which distributions are being compared. For example, if h1 data points are twice as likely as h2 data points to become ambiguous going from the degree-1 to the degree-0 distribution, then I assume h1 data points are twice as likely as h2 data points to become ambiguous going from the underlying distribution to the degree-1 distribution.

(31) Ambiguity Loss Ratio Assumption

$$\mathbf{Lutod1 = Ld1tod0} = \frac{\mathbf{u1d1'\ -\ u1d0}}{\mathbf{u2d1'\ -\ u2d0}}$$

Now, we have all the pieces in place to write an equation that relates the ambiguity loss of h1 data points to the ambiguity loss of h2 data points going from the underlying distribution to the degree-1 distribution. The intuition is laid out in (32).

(32) Intuition to relate ambiguity loss from underlying to degree-1 distribution

% of h1 data points “lost” = **Lutod1*** % of h2 data points “lost”

$$\frac{\#\text{ of h1 data points lost}}{\text{total }\#\text{ of h1 data points}} = \mathbf{Lutod1} * \frac{\#\text{ of h2 data points lost}}{\text{total }\#\text{ of h2 data points}}$$

This intuition can be instantiated as in (33). We can then use the equations we have already derived to solve for γ, the probability of accessing h1 in the underlying distribution.

(33) Solving for γ

(from (32)) $\dfrac{a1utod1}{u1} = \mathbf{Lutod1} * \dfrac{a2utod1}{u2}$

(from (31)) $\dfrac{a1utod1}{u1} = \mathbf{Ld1tod0} * \dfrac{a2utod1}{u2}$

(from (25)) $\dfrac{a1utod1}{\gamma * \mathbf{d0}} = \mathbf{Ld1tod0} * \dfrac{a2utod1}{u2}$

(from (27)) $\dfrac{a1utod1}{\gamma * \mathbf{d0}} = \mathbf{Ld1tod0} * \dfrac{a2utod1}{\mathbf{u2d1'} + \mathbf{ad1'} - a1d1}$

(from (30)) $\dfrac{a1d1}{\gamma * \mathbf{d0}} = \mathbf{Ld1tod0} * \dfrac{\mathbf{ad1'} - a1d1}{\mathbf{u2d1'} + \mathbf{ad1'} - a1d1}$

(from (25)) $\dfrac{\gamma * \mathbf{d0} - \mathbf{u1d1'}}{\gamma * \mathbf{d0}} = \mathbf{Ld1tod0} * \dfrac{\mathbf{ad1'} - (\gamma * \mathbf{d0} - \mathbf{u1d1'})}{\mathbf{u2d1'} + \mathbf{ad1'} - (\gamma * \mathbf{d0} - \mathbf{u1d1'})}$

$$\gamma^2(\mathbf{Ld1tod0} + 1)(\mathbf{d0}^2)$$
$$+ \gamma(\mathbf{d0})(\mathbf{d0} + \mathbf{u1d1'} - \mathbf{Ld1tod0} * (\mathbf{ad1'} + \mathbf{u1d1'}))$$
$$+ (\mathbf{-1})(\mathbf{d0} * \mathbf{u1d1'}) = 0$$

Now, we can use the quadratic formula to solve for γ.

$a = (\mathbf{Ld1tod0} + 1)(\mathbf{d0}^2)$

$b = (\mathbf{d0})(\mathbf{d0} + \mathbf{u1d1'} - \mathbf{Ld1tod0} * (\mathbf{ad1'} + \mathbf{u1d1'}))$

$\mathrm{c} = (\mathbf{-1})(\mathbf{d0} * \mathbf{u1d1'})$

$$\gamma = \frac{-(\mathbf{d0})(\mathbf{d0} + \mathbf{u1d1'} - \mathbf{Ld1tod0} * (\mathbf{ad1'} + \mathbf{u1d1'}))}{2(\mathbf{Ld1tod0} + 1)(\mathbf{d0}^2)}$$

$$+/- \frac{\sqrt{((\mathbf{d0})(\mathbf{d0} + \mathbf{u1d1'} - \mathbf{Ld1tod0} * (\mathbf{ad1'} + \mathbf{u1d1'})))^2 - 4(\mathbf{Ld1tod0} + 1)(\mathbf{d0}^2)((\mathbf{-1})(\mathbf{d0} * \mathbf{u1d1'}))}}{2(\mathbf{Ld1tod0} + 1)(\mathbf{d0}^2)}$$

This formula can be easily resolved once we insert the observable quantities from the historical corpus, as we will see in the next section. Once we have solved for γ, we have the probability with which h1 is accessed in the underlying distribution. We can calculate the probability with which h2 is accessed in the underlying distribution by using 1 - γ.

4.4.2.2.2 A Concrete Example of Deriving p_{VO} from Historical Data

For our Old English corpus, let h1 be the OV word order hypothesis and h2 be the VO word order hypothesis. I will step through the derivation of the underlying p_{VO} value at 1000 A.D. First, we observe the various quantities available from the historical corpus at 1000 A.D.

	Unamb OV (h1)	Unamb VO (h2)	Amb
Degree-0	**1389**	**936**	**7480**
Degree-1	**3844**	**1583**	**2132**
Underlying Distribution	*u1*	*u2*	0

Table 4.6. Quantities available from historical corpora and quantities to derive. Quantities in **bold** are gathered from historical corpora. Quantities in *italics* must be derived and are used to calculate the average p_{VO} in the population.

Then, we normalize the degree-1 quantities to the degree-0 quantities. The total quantity of degree-0 data **d0** is 1389 + 936 + 7480 = **9805**. The total quantity of degree-1 data **d1** is 3844 + 1583 + 2132 = **7559**. To normalize the degree-1 quantities, we therefore multiply each quantity by **(d0/d1)** = **(9805/7559)**.

	Unamb OV (h1)	Unamb VO (h2)	Amb
Degree-0	**1389**	**936**	**7480**
Degree-1	**4986**	**2053**	**2766**
Underlying Distribution	*u1*	*u2*	0

Table 4.7. Data quantities after normalization.

We then calculate the ambiguity loss ratio between the degree-1 and degree-0 distribution, **Ld1tod0**.

$$\text{(34) } \mathbf{Ld1tod0} = \frac{\mathbf{u1d1' \ - \ u1d0}}{\mathbf{u2d1' \ - \ u2d0}} = \frac{\mathbf{4986 \ - \ 1389}}{\mathbf{2053 \ - \ 936}} \approx 3.22$$

So, we see that the OV data points are over three times as likely to become ambiguous as the VO data points at 1000 A.D. I assume that this loss ratio is the same going from the underlying distribution to the degree-1 distribution (**Lutod1**), that is that OV data points are three times as likely as VO data points to become ambiguous going from the underlying to the degree-1 distribution.

We now have all the quantities we need to calculate γ (from (33)).

$$\gamma = \frac{-(\mathbf{9805})(\mathbf{9805 \ + \ 4986 \ - \ 3.22} * (\mathbf{2766 \ + \ 4986}))}{2(\mathbf{3.22} \ +1)(\mathbf{9805}^2)}$$

$$+/- \frac{\sqrt{((\mathbf{9805})(\mathbf{9805 \ + \ 4986 \ - \ 3.22} * (\mathbf{2766 \ + \ 4986})))^2 - 4(\mathbf{3.22} \ +1)(\mathbf{9805}^2)((\mathbf{-1})(\mathbf{d0} * \mathbf{4986'}))}}{2(\mathbf{3.22} \ +1)(\mathbf{9805}^2)}$$

Solving for γ, we obtain 0.766 and -.299. Since we know γ is a probability and must be between 0.0 and 1.0, the correct solution for γ is 0.766. So, given these historical data distributions, I estimate that the OV word order option was accessed with probability 0.766 at 1000 A.D. The VO word order option was thus accessed with probability 1-0.766 = 0.234. Since we are tracking the probability with which the VO word order option is accessed, p_{VO} is 0.234 at 1000 A.D. The average p_{VO} values in the population at the other two periods of time we consider (1000-1150 A.D. and 1200 A.D.) can be calculated in the same fashion by using the quantities in table 4.1. Table 4.8 below displays the three p_{VO} values I will be using in my simulations.

	Degree-0 Clauses			Degree-1 Clauses			Underlying
	Total	OV Unamb	VO Unamb	Total	OV Unamb	VO Unamb	p_{VO}
1000 A.D.	9805	1389	936	7559	3844	1583	**.234**
1000 – 1150 A.D.	6214	624	590	3636	1759	975	**.310**
1200 A.D.	1282	180	190	2236	551	1460	**.747**

Table 4.8. Data from historical corpora and calculated p_{VO}.

To model the data from the historical corpus, a population must start with an average p_{VO} of 0.234 at 1000 A.D., reach an average p_{VO} of 0.310 between 1000 and 1150 A.D.[30], and reach an average p_{VO} of 0.747 by 1200 A.D.

[30] This is what is meant by calibration. If the population is unable to reach this checkpoint, it is unfair to compare its p_{VO} at 1200 A.D. against other populations' p_{VO} values at 1200 A.D. The value which must be calibrated is the learner's confidence value *c* in the current piece of data, which determines how much the current p_{VO} is updated for a given data point.

4.4.2.3 Answering Questions About Learning Filters

Armed with these models of individual-level learning and population-level change, we can now answer two questions about filters on the learner's intake. First, I address the question of descriptive *sufficiency*: can an Old English population whose learners filter their intake down to the degree-0 unambiguous data shift from a strongly OV biased distribution to a strongly VO biased distribution at the appropriate time? Recall that the data intake set is significantly smaller than the data input set, and so there is a potential data sparseness problem. Moreover, recall that exactly the right amount of misconvergence on the p_{VO} value must happen for each set of new population members in order for the population as a whole to change at the correct rate. We can ask if input filtering in the specified manner can cause this precise amount of misconvergence.

Second, we address the question of *necessity*. If the proposed intake filtering is sufficient to cause an Old English population to change at the correct rate, is it in fact necessary? One might wonder if an Old English population that does not use either filter, or only uses one (either unambiguous data or degree-0), would achieve the same results. With the model described here, we can find out.

4.5 Modeling Results

4.5.1 Sufficient Filtering

We first examine the descriptive sufficiency of the data intake filters. Does our simulated Old English population, whose learners filter their intake down to the degree-0 unambiguous data, undergo change at the historically attested rate? Figure 29 shows the average population p_{VO} over time. Based on these simulation data, an Old English population using these filters can indeed shift from a strongly OV-biased distribution to a strongly VO-biased distribution at the historically correct time. Specifically, these filters yield a data set with the right bias at each point in time. This then allows individual learners in the population to misconverge exactly the right amount, which then leads to population-level change at the correct rate.

Moreover, we can see that the concern over data sparseness can be put aside. Despite the significantly smaller quantity of data that comprises the intake for these learners, the trajectory of the population is still in line with the known historical trajectory. We also note that the S-shaped curve so often observed in language change (Bailey, 1973; Weinreich, Labov, & Herzog, 1968; Osgood & Sebeok, 1954; among others) emerges here from the learners filtering their input and the subsequent small changes spreading through an exponentially growing population.[31]

[31] As mentioned previously, this demonstrates that external factors are not *necessary* to cause swift population-level change. Here, the population-level change results from internal factors: the language-learning mechanism at the individual-learning level.

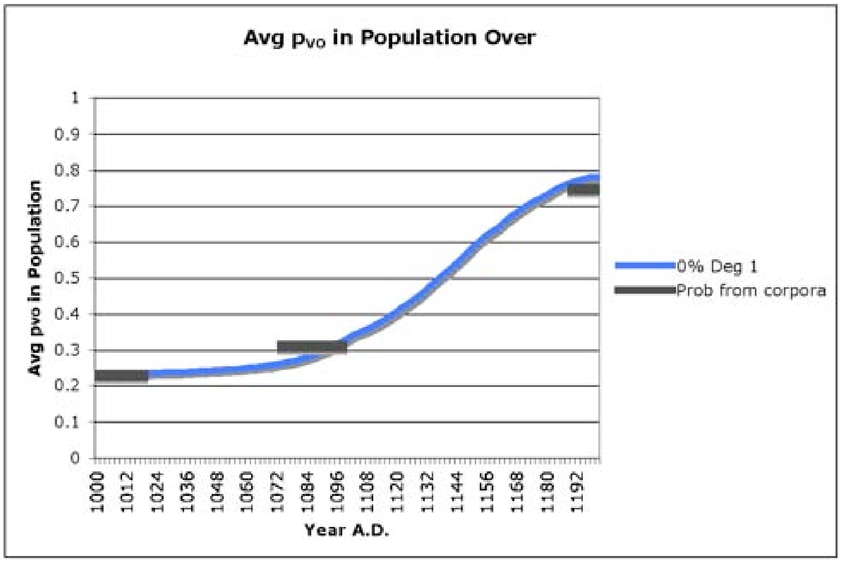

Figure 29. The trajectory of a population learning only from degree-0 unambiguous data, compared against estimates from historical corpora.

4.5.2 Necessary Filtering

We have just seen that these data intake filters are sufficient to cause the right rate of population-level change to occur. But are they necessary? Specifically, we wish to know if language change can occur at the historically attested rate without these filters. I examine the effects of removing each filter in turn, and then the effects of removing both.

4.5.2.1 Removing the Unambiguous Data Filter

I examine the unambiguous data filter first. A model could reasonably choose to drop this filter and assume that a learner attempts to activate the update algorithm for data that are ambiguous. In particular, the learner then requires some strategy to extract information from a given ambiguous data point. One simple strategy is for the learner to have a preference for analyzing strings as base-generated. This strategy would cause the learner to discard any analyses involving movement (for example, V2 movement) until forced to do so (Fodor, 1998b).

The effect of this strategy for the OV/VO word order cases we consider in Old English is that many more data points are used by the learner. Primary among these new data points are those of the form *Subject TensedVerb Object*. When V2 movement was considered in the analysis, this was ambiguous between OV order (OV, +V2) and VO order (VO, +/-V2), as we saw in example (13). However, if non-movement analyses are given preference, then the learner would take this ambiguous data point as evidence in favor of the VO word order hypothesis. Table 4.9 displays the data intake distribution for a learner who does not use an unambiguous data filter, as well as the OV word order bias at different points in time.

	D0 Total # Clauses	OV Data Intake	VO Data Intake	D0 OV Bias w.r.t. the input[a]	D0 OV Bias w.r.t. the intake[b]
1000 A.D.	9805	2537	3889	-13.8%	**-21.0%**[c]
1000 – 1150 A.D	6214	1221	2118	-14.4%	**-26.9%**
1200 A.D.	1282	389	606	-16.9%	**-21.8%**

Table 4.9. OV order bias in the degree-0 (D0) clauses. [a] We derive the bias for the OV order with respect to the *input* by subtracting the quantity of VO data from the quantity of OV data, and then dividing by the total number of data points in the input. For instance, the D0 OV bias at 1000 A.D. is calculated as (2537-3889)/9805 = 13.8%. [b] We derive the bias for the OV order with respect to the *intake* by subtracting the quantity of VO data from the quantity of OV data, and then dividing by the total number of data points in the intake. For instance, the D0 OV bias at 1000 A.D. is calculated as (2537-3889)/(2537+3889) = 21.0%. [c] Note that a negative OV bias means that the distribution is VO-biased.

A very serious problem becomes apparent: even at the earliest time period when the population is supposed to be strongly OV-biased, the data intake distribution strongly favors the VO order. The VO word order has a 21.0% bias in the data intake at 1000 A.D (and a 13.8% bias in the input). Thus, about 21 out of every 100 data points encountered in the intake are biasing the learner towards the VO hypothesis. A population of learners using this data intake distribution could not remain strongly OV-biased for very long, and certainly not until 1150 A.D.

Therefore, I conclude that dropping the unambiguous data filter in this way will not allow the model to simulate what is actually observed in the Old English population. So, these results suggest that the unambiguous data filter is necessary.[32]

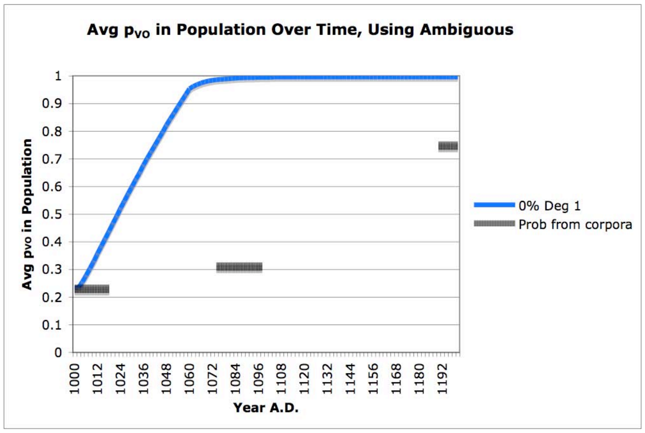

Figure 30. The trajectory of a population learning only from degree-0 data (ambiguous and unambiguous), compared against estimates from historical corpora.

[32] Unless we can find a strategy to deal with ambiguous data which includes a different set of data as intake, or values the ambiguous data in a manner that gives the OV hypothesis the advantage early on. The strategy explored here was the simplest (and most justifiable) one I could devise, but there may be more complex strategies that yield the desired results. If so, then we would need an explanation for the learner's knowledge and adoption of these more complex strategies.

4.5.2.2 Removing the Degree-0 Filter

I turn now to the degree-0 data filter. Suppose we drop this filter and allow the modeled learner to activate the update algorithm for both matrix (degree-0) and embedded (degree-1) clauses. Note that this learner still has the unambiguous data filter, and so will only activate the update procedure if the learner perceives the data point as unambiguous. Recall from table 4.1 that the degree-1 data intake distribution has a much higher OV bias before 1150 A.D. (28.7 – 41.7%). Given how high this OV bias is, it is possible that if there were enough degree-1 data in the input set, the learner would converge on a final p_{VO} that is too OV-biased. This slows the rate of change from OV-biased to VO-biased, and so a population made up of such learners would proceed much more slowly towards becoming VO-biased. I have estimated from the historical record that the Old English population should have an average p_{VO} value of 0.747 at 1200 A.D. This is the mark a simulated population must then reach.

With the model presented here, we can test the population-level effects of different compositions of data in the input set of the individual learner. Specifically, we can see how much (strongly OV-biased) degree-1 data can be in the input (and thus in this learner's intake) and still have the population as a whole be VO-biased enough by 1200 A.D. We can then compare this threshold against the estimated amount of degree-1 data available to learners and see if the degree-0 data filter is necessary. If the estimated amount of degree-1 data available to learners is less than the permissible threshold that allows correct population-level behavior, then the degree-0 filter is not necessary. The same population-level results can be obtained with or without the filter. In contrast, if the estimated amount of degree-1 data available to learners is greater than the permissible threshold, then we have support for the necessity of the degree-0 filter. This is because only by ignoring the degree-1 data available in the input can correct population-level behavior be obtained.

Figure 31 displays the average p_{VO} in the population at 1200 A.D. for 6 Old English populations whose learners had their input composed of different percentages of degree-1 data. For these populations, all the degree-1 data was in the intake set. Thus, a population with 16% degree-1 data in the input set activated the updating procedure for the 84% of the unambiguous data points that were degree-0 and the 16% of the unambiguous data points that were degree-1. Data points that were ambiguous were ignored.

The modeling results suggest that having even 4% degree-1 data available in the input (and thus in the learner's intake) is enough to prevent the simulated Old English population from reaching an average p_{VO} of 0.747 by 1200 A.D. We must now compare this threshold to the estimated amount of degree-1 data in the input to Old English learners.

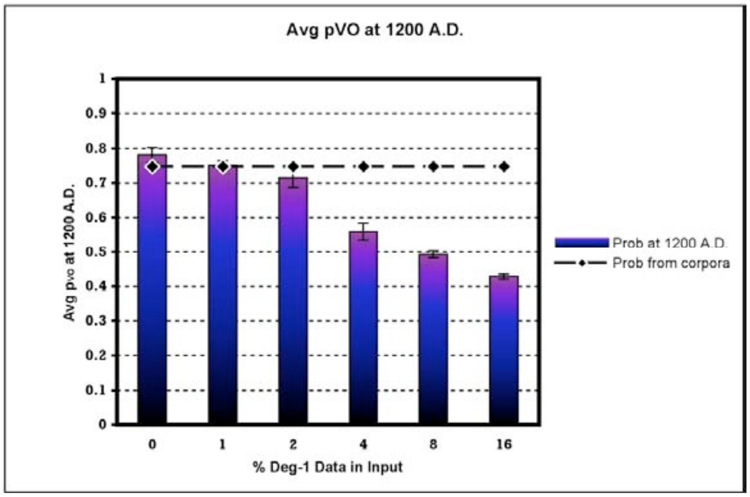

Figure 31. Average probability of using VO order at 1200 A.D. for populations with differing amounts of degree-1 data available during learning, as compared to the estimated average from historical corpora. Confidence intervals of 95% are shown.

I assume that amount of degree-1 child-directed data is approximately the same no matter what the time period (and I am currently unaware of studies that suggest otherwise). Given this, we can examine samples of modern English child-directed data to see what its composition is. The two samples I chose were a portion of the CHILDES database (MacWhinney, 2000) and some young children's stories (some of which can be found at http://www.magickeys.com/books/index.html). I used CHILDES since it is recorded speech to children and young children's stories because it is (storytelling) language designed to be directed at children. As we can see from Table 4.10, the CHILDES sample has approximately 8.8% degree-1 data points while the young children's stories sample has approximately 23.9% degree-1 data points. I take the average of these two sources to get an estimate of about 16% degree-1 data available in children's input. This is very similar to the 15% degree-1 data estimate from Sakas (2003), who examined several thousand sentences from the CHILDES database.

The modeling results (see figure 31) show that input comprised of 16% degree-1 data causes the simulated Old English population to be far too slow in shifting to a strongly VO-biased distribution. This is much higher than the permissible threshold of approximately 2%. Unless there is a way for the learner to allow in only an eighth of the degree-1 data available in the input, these results suggest that the degree-0 data intake filter is also necessary.[33]

[33] Another option is for the learner to weight the degree-1 data's influence so it is only an eighth as strong as the degree-0's influence. This particular weighting would then have to be justified.

A subsection of CHILDES				
Total Utterances	Total Data Points[a]	Total D0	Total D1	% D1
4068	2760	2516	244	8.8
Sample D0 Utterances “What’s that?”, “I don’t know.”, “There’s a table.”, “Can you climb the ladder?”, “Shall we stack these?”, “That’s right.”		Sample D1 Utterances “I think it’s time…”, “Look what happened!”, “I think there may be one missing.”, “Show me how you play with that.”, “See if you can get it.”, “That’s what he says.”		

Young Children’s Stories				
Total Utterances	Total Data Points[a]	Total D0	Total D1	% D1
4031	3778	2955	927	23.9
Sample D0 Utterances “Ollie is an eel.”, “She giggled.”, “…but he climbs the tree!”, “This box is too wide.”, “…to gather their nectar.”[b], “This is the number six.”		Sample D1 Utterances “…that even though he wishes hard,…”, “…that only special birds can do.”, “…that can repeat words people say.”, “…when the sun shines.”, “…that goes NEIGH…NEIGH…”, “…know what it is?”		

Table 4.10. Data gathered from speech directed to young children. [a] The number of data points is much less than the number of utterances since many of these utterances include “Huh?” and exclamations like “A ladder!” in the case of the spoken CHILDES corpus. For the young children’s stories, there are often “sentences” like “Phew!” and “Red and yellow and green” which were excluded under Total Data Points. [b] I note that clauses with infinitives such as “…to gather their nectar” are included under degree-0 data, based on Lightfoot’s (1991) definition of clause-union structures as degree-0. If this were not the case, the percentage of degree-1 clauses would only be higher than what I have calculated here – thus, this is a lower bound on the amount of degree-1 data available in the input.

4.5.2.3 Removing Both Filters

We have just observed that the loss of each of the data intake filters has a different effect on the rate of change at the population-level. Without the unambiguous data filter, the intake distribution is too heavily VO-biased. The population becomes strongly VO-biased too soon, and so changes too quickly. Without the degree-0 data filter, the intake distribution is too heavily OV-biased. The population becomes strongly VO-biased too late, and so changes too slowly. Given these opposite effects, one might wonder if dropping both filters would allow the simulated population to change at the correct rate. We must again examine the data intake distributions that learners would be using to see the effects of removing both filters.

	Total # Clauses	OV Data Intake	VO Data Intake	D0 OV Bias w.r.t. the input[a]	D0 OV Bias w.r.t. the intake[b]
Degree-0 Data	9805	2537	3889	**-13.8%**	**-21.0%**[c]
Degree-1 Data	7559	4650	2610	26.9%	28.1%

Table 4.11. OV order bias at 1000 A.D. with no filters. [a] We derive the bias for the OV order with respect to the *input* by subtracting the quantity of VO data from the quantity of OV data, and then dividing by the total number of data points in the input. For instance, the D0 OV bias at 1000 A.D. is calculated as (2537-3889)/9805 = 13.8%. [b] We derive the bias for the OV order with respect to the *intake* by

subtracting the quantity of VO data from the quantity of OV data, and then dividing by the total number of data points in the intake. For instance, the D0 OV bias at 1000 A.D. is calculated as (2537-3889)/(2537+3889) = 21.0%. [c] Note that a negative OV bias means that the distribution is VO-biased.

In order for the Old English population to remain strongly OV-biased before 1150 A.D., the data intake distribution must at least be OV-biased at 1000 A.D. As we can see from table 4.11, the degree-0 data intake is heavily VO-biased (21.0% VO data bias). In order to drop the VO bias in the intake down to zero (so the OV order has at least a fighting chance with learners at 1000 A.D.), about 43% of the intake would need to consist of degree-1 data.

My estimate of the available amount of degree-1 data in child-directed data suggests that *less than half* of this amount of degree-1 data is available, at best (16%).[34] So, I conclude that we cannot drop both the unambiguous data filter and the degree-0 data filter, lest the population be driven to become strongly VO-biased too soon. The claim that both data intake filters are necessary is thus strengthened.

4.6 General Discussion

4.6.1 Necessary Filters

The results presented here serve as an existence proof that a population model whose individual learners employ data intake filtering can handle the specific case of word order change in Old English. The two critical filters are (a) use only data perceived as unambiguous and (b) use only degree-0 data. This means that the update procedure is only activated when data points obeying these constraints are encountered. Otherwise, the update procedure is not activated and the data points are effectively ignored for the purposes of learning.

I now examine what effects input filtering in general could have on language change, as well as the feasibility of input filtering.

4.6.2 Intake Filtering and Language Change

The nature of the input filter may be what differentiates situations of language change from situations of stable variation. If the intake becomes too mixed for the child to converge on the same probability weighting as the adult, then language change will occur. In cases where only one structural option is used in the adult population (as is often the case), the adult probability distribution will be 0.0 or 1.0. Given children's tendency to generalize to an extreme value from noisy data (Hudson Kam & Newport, 2005), the intake would have to be quite mixed in order to force children away from the adult distribution.

In this way, we see that learning can tolerate some variation in the input without causing the language to change. In this, our model's behavior differs notably

[34] Moreover, since not all the data in the input becomes intake, even more than 43% of the input would need to consist of degree-1 data. Give that, the available quantity of degree-1 data is certainly insufficient.

from Briscoe's (2000), who observed constant oscillation in the population due to slight variation in the input to learners. The model here differs from his by using only unambiguous data to update the learner's hypothesis. I also allow the learner's final probability to be a value other than 0.0 or 1.0. I hypothesize that this is what yields the historically correct behavior. In addition, the model here has more realistic estimates for input quantity, population size, and learner lifespan.

4.6.3 The Feasibility of Filters

One might well be skeptical of the generality of the proposed filters. The unambiguous data filter in particular raises the question of how abundant such data points are for any given learning problem and the complexity of determining if a given data point is unambiguous. As a concrete example of both these issues for the word order case considered here, we can look to the "cartographic" approach to syntax (Rizzi, 2004; 1997; Cinque, 1999). This approach suggests that there are several positions in front of the VP that the Verb can move to if V2 movement is used. Languages are thought to differ on exactly which position it is. Given that, even knowing V2 movement has happened does not allow an unambiguous analysis of the sentence with respect to V2 movement; the learner still has more than one option for the Verb's exact position. If the initial intake is to contain any data points at all, it may be necessary to allow data points that are actually ambiguous to be *perceived* as unambiguous at the initial stages of learning.

If the learner is using cues to identify unambiguous data, then the level of specificity for a cue may be abstract enough to perceive ambiguous data as unambiguous. For instance, a cue may only specify one general position in front of the VP to identify V2 movement, rather than the multiple positions that the cartographic approach advocates. Only later would the learner then elaborate cues to include multiple positions in front of the VP. If the learner is using parsing to identify unambiguous data, then the learner could initially use a subset of the set of parameters an adult would use when parsing.[35] Later on, when more parameter values are known, the learner would expand the set of parameters used for parsing.

Another approach for both cues and parsing is that the learner has default values or assumptions (Fodor, 1998b) that are in place until the learner is forced to the marked values or assumptions. For example, in the word order case discussed here, the learner might assume as a default that there is no movement (thus perceiving simple SVO structures as unambiguous for VO word order). This assumption would then need to be revised at a later stage. The cost of reanalysis may not trivial, however, particularly when parameters and assumptions interact with each other.

Suppose, for instance, that default assumption A1 (e.g. no movement) allows the learner to perceive "unambiguous" data for a given value of P1 (e.g. OV/VO order), say, P1a (e.g. VO order). Later on, the learner is forced to remove default assumption A1. Suppose the lack of assumption A1 causes the learner to observe that (a) the "unambiguous" data for P1a are now ambiguous (e.g. SVO data) and (b) there now exist "unambiguous" data for P1b (e.g. more OV data). The learner must now

[35] A candidate set for the initial pool of parameters might be derived from a hierarchy of parameters, along the lines of the one based on cross-linguistic comparison that is described in Baker (2005, 2001).

re-evaluate the correct value for parameter P1 (OV/VO order), and so is delayed in attaining the adult target state. This same situation occurs when there are multiple parameters interacting (say, +/- V2 movement and OV/VO order). The issue of identifying unambiguous data in a system with multiple interacting parameters will be discussed in the next chapter.

The identification of unambiguous data is significantly aided by the assumption that parameters are independent structural pieces. Suppose we assume *n* parameters with 2 options each. If all parameters are independent, then every data point has at most 2*n* possible structural pieces that can be used to analyze it (Fodor, 1998a; 1998b; Sakas & Fodor, 1998). In contrast, if parameters are not independent, every data point can be analyzed with 2^n possible structures (since each "structure" is a combination of the smaller 2*n* structural pieces). It is thus enormously more efficient for ambiguity analysis to have independent parameters.

Moreover, if parameters are independent, data are unambiguous relative to a particular parameter. A given data point may be unambiguous for parameter P1 (e.g. OV ordering) while being ambiguous for many other parameters (e.g. wh-fronting). In contrast, if parameters are not independent, only data points that are unambiguous for *all* parameters are perceived as unambiguous – for otherwise, more than one structure of the available 2^n structural pieces leads to a successful analysis. Such data points are likely to be extremely sparse, if they exist at all.

4.6.4 Future Directions

Despite the ground covered in this chapter, there are of course a number of avenues that remain to be explored. The first concerns the relaxation of the unambiguous data filter, the second concerns the implementation of population models, and the third concerns experimental extensions.

In section 4.5.2.1, I explored one principled way a learner might use ambiguous data, which was to ignore possible movement rules in the system and assume that surface word order matched the underlying word order of the system. So, the hypothesis consistent with the surface order was fully credited for those data points, i.e. a data point with *Verb Object* anywhere in it would be credited to the Verb Object hypothesis. But there are other strategies that a learner might employ when encountering ambiguous data.

One method is to weight ambiguous data points such that they're not as influential as unambiguous data. In fact, I instantiated a method to do precisely this in the case study of anaphoric *one* in chapter 3, and the actual instantiation appears in the update procedure. The same concept of weighting could be applied to the syntax case examined in this chapter. If learners weight ambiguous data less than unambiguous data, it may be possible for them to achieve successful acquisition. If so, it behooves us to know what the successful weightings are for ambiguous and unambiguous data – and if we can find any experimental evidence to support such weightings.

Continuing the idea of weighting data, models of populations (such as the one examined here) can include additional sociolinguistic complexity in the relationships of the speakers that impact how learners view the data. Learners might, for instance, be more influenced by speakers who are in close spatial proximity, have a kinship relationship, or are from the same or higher social class. This weighting again would be instantiated in the update procedure. In addition, the frequency of various data types in the data intake distribution could depend on what speakers are nearby and/or are prominent in the learner's life. Family members will be a more frequent source of data than random, spatially distant population members.

Finally, the existence of data intake filtering for learning syntax – and specifically, using data perceived as unambiguous – can be explored in experimental regimes such as artificial language experiments for both adults (Thompson & Newport, 2007; Bonatti et al., 2005; Newport & Aslin, 2004) and children (Saffran, Aslin, & Newport, 1996; Saffran, Newport, & Aslin, 1996; among others). Specifically, learners could be exposed to data that would favor one word order if ambiguous data is used, but favor the other order if only unambiguous data is used. The generalization learners extract from such a dataset would implicate what data they use for learning.

4.6.5 Conclusion

In this chapter, I have investigated the effect of data intake filters in a system where the target adult state is a probability distribution between two opposing options for a single parameter. This was accomplished by employing language change modeling and using the assumption that a given case of language change was driven by language learning. Specifically, I adopted Lightfoot's (1991) assumption that Old English language change was driven by imperfect learning. The goal of the modeling was to see if I could replicate the precise amount of imperfect learning that causes the Old English population to change at a certain rate. As we have seen, this imperfect learning can result when the two data intake filters of unambiguous data and degree-0 data are used. Moreover, the historically correct population-level behavior does not result when either or both of the two filters is discarded, primarily because the data intake then does not have the proper bias in its distribution. Thus, through the language change model, I have provided empirical support for data intake filtering in language learning.

Now that we have seen evidence for the necessity of data intake filtering, we can now explore the feasibility of data intake filtering. This is particularly important for the unambiguous data filter, since identifying unambiguous data is a nontrivial task. In fact, it is quite reasonable to wonder how to identify unambiguous data in a system more complex than the one I have considered in this chapter (which considered 2 interacting parameters: OV/VO word order and +/- V2 movement). In the next chapter, I will examine the feasibility of the unambiguous data filter in a more complex system with 9 interacting parameters: English metrical phonology (Dresher, 1999).